A MODEL ENSEMBLE APPROACH FOR AUDIO-VISUAL SCENE CLASSIFICATION

Technical Report

Qing Wang[1], *Siyuan Zheng*[1], *Yunqing Li*[1], *Yajian Wang*[1], *Yuzhong Wu*[2,4], *Hu Hu*[3], *Chao-Han Huck Yang*[3], *Sabato Marco Siniscalchi*[3,5], *Yannan Wang*[2], *Jun Du*[1], *Chin-Hui Lee*[3],

[1] NELSLIP, University of Science and Technology of China, Hefei, China
[2] Tencent Ethereal Audio Lab, Tencent Corporation, China
[3] School of Electrical and Computer Engineering, Georgia Institute of Technology, GA, USA
[4] DSP & Speech Technology Laboratory, The Chinese University of Hong Kong, Hong Kong
[5] Computer Engineering School,University of Enna Kore, Italy

ABSTRACT

In this technical report, we present our approach to Task 1b - Audio-Visual Scene Classification (AVSC) in the DCASE 2021 Challenge. We employ pre-trained networks trained on image datasets to extract video embedding whereas for audio embedding models trained from scratch are more appropriate for feature extraction. We propose several models for the AVSC task based on different audio and video embeddings using early fusion strategy. Besides, we propose to use acoustic and visual segment model (AVSM) to extract text embedding. Data augmentation methods are used during training. Furthermore, a two-stage classification strategy is adopted by leveraging on score fusion of two classifiers. Finally, model ensemble of two-stage AVSC classifiers is used to obtain more robust predictions. The proposed systems are evaluated on the development dataset of TAU Urban Audio Visual Scenes 2021. Compared with the official baseline system, our approach can achieve a much lower log loss of 0.141 and a much higher accuracy of 95.3%.

Index Terms— Audio-visual scene classification, audio embedding, video embedding, text embedding, acoustic and visual segment model, data augmentation, model ensemble

1. INTRODUCTION

Acoustic scene classification (ASC) aims to classify audio recordings into pre-defined environment classes, which has become an important task of the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge in recent years [1, 2]. The ASC task has attracted more and more researchers and state-of-the-art results are achieved by deep-learning based methods, such as convolution neural networks (CNNs) [3, 4, 5].

Different from the ASC task in DCASE 2020 Challenge, an audio-visual scene classification (AVSC) task is introduced for the first time in DCASE 2021 Challenge. Besides audio recordings, the AVSC task provides video clips which contain more intuitive content. The AVSC task is focused on scene classification using audio and video modalities. The key issue is to learn effective audio and video embeddings which are appropriate for scene classification.

In this technical report, we present our approach for the AVSC task. Models pre-trained on two image databases (ImageNet [6] and Places365 [7]), such as DenseNet [8], ResNet [9], ResNeSt [10], and VGG [11] are adopted for video embedding learning. For audio embedding, we investigate pre-trained models such as VGGish [12] and PANN [13] which are both trained on AudioSet dataset [14] together with models trained from scratch to extract audio embedding. Besides audio and video embeddings, we propose to use acoustic and visual segment model (AVSM) for text embedding extraction. AVSM model is able to analyze the scene information in detail, which may be complimentary to audio and video embeddings. Audio, video and text embeddings are concatenated by using early fusion strategy before fed to the classifier. Based on our ASC solution in DCASE 2020 Challenge [4], two-stage classification procedure helps improve the accuracy, so we also build a two-stage AVSC system in this task which includes a three-class classifier and a ten-class classifier. We train several models with different audio, video and text embeddings. Moreover, data augmentation methods are used to reduce the overfitting problem. Model ensemble of different systems is adopted to provide a more robust scene prediction.

The rest of the technical report is organized as follows. Section 2 presents our proposed approach in detail, including data augmentation, feature embedding and two-stage classification. Section 3 describes experimental setup and evaluation results on the development dataset. Finally, conclusions are summarized in Section 4.

2. PROPOSED APPROACH

In the proposed approach, several deep neural network (DNN) models are trained for the AVSC task. Figure. 1 shows our proposed audio-visual and textual scene classification (AVTSC) model when adopting audio, video and text embeddings simultaneously. If the text embedding module as shown in the black dashed rectangular box in Figure. 1 is removed, we regard it as the AVSC model. We will describe the three parts of our proposed approach in detail: the data augmentation, the feature embedding and the two-stage classification.

2.1. Data Augmentation

In our previous work [15], data augmentation methods are proven to be effective in ASC task of DCASE 2020 Challenge. Thus we adopt several data augmentation methods in our approach to improve the generalization abilities of DNN models. These data augmentation methods are listed below:

- SpecAugment: It is a simple data augmentation method and was first proposed for automatic speech recognition (ASR)

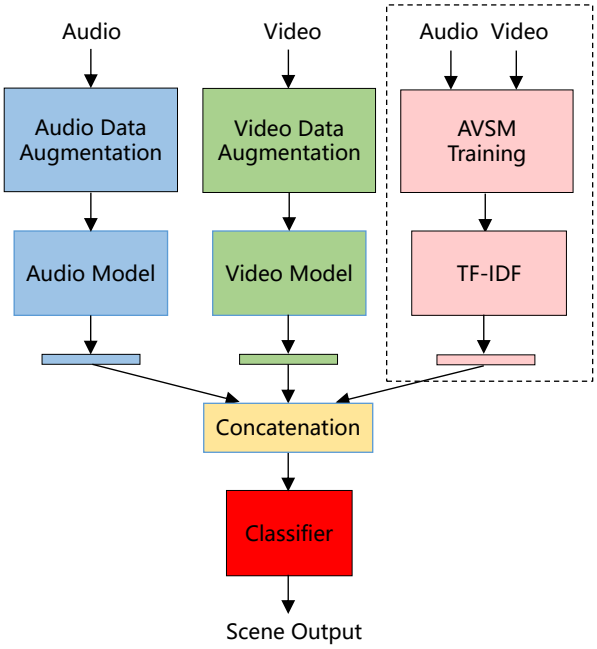

Figure 1: The proposed AVTSC model for audio-visual scene classification.

[16]. We didn't do time warping in this study. For time and frequency masking, we set the masking parameter to 10% of the dimensions. It is applied in log-mel features.

- Channel confusion: Two channels of input audio feature are randomly swapped.
- Pitch shifting: It is performed on audio waveforms to randomly shift the pitch based on the uniform distribution. It is only used in training audio models.
- Speech changing: It is performed on audio waveforms to randomly change the speech of audio recordings. It is only used in training audio models.
- Noise adding: It is performed on audio waveforms to add random Gaussian noise. It is only used in training audio models.
- Audio mixing: Two audio recordings from the same scene class are mixed to generate a new sample with the same label. It is only used in training audio models.
- Mixup: It was proposed in [17] and the parameter alpha is set to 0.4. Two batches of data are randomly mixed in each step along with the corresponding labels when training audio models. And 20% of the training data are employed with mixup when building the multimodal systems.
- AutoAugment: It was proposed in [18] for image recognition. We adopt three sub-policies in the search space, namely Sharpness, Contrast and IdentityMapping. For each image, two operations are randomly selected to be applied in sequence.

2.2. Feature Embedding

Wang *et al.* [19] proposed the baseline system for the AVSC task of DCASE 2021 Challenge based on the OpenL3 network [20]. We investigate several pre-trained models for audio and video embedding learning. Besides, AVSM approach is used to provide textual information through latent semantic analysis.

2.2.1. Audio Embedding

The VGGish [12] and PANN [13] networks are both trained on AudioSet [14]. We use these two pre-trained models to extract audio embeddings and apply them in this AVSC task by transfer learning. The audio embedding dimensions of VGGish and PANN are 128 and 2048, respectively. In our previous works [4, 15], CNN-based models achieved promising performance in acoustic scene classification. To better leverage the acoustic presentation in these models, we propose to adopt models trained using the audio data of this task, namely FCNN and Resnet [4], to extract high-level embedded features which may contain more useful acoustic information. And the embedding dimensions of FCNN and Resnet are 160 and 1280, respectively.

2.2.2. Video Embedding

ImageNet is a large-scale hierarchical image database [6], which is widely adopted for image classification. Various DNN architectures are trained on ImageNet to solve the image classification task, for example, VGG [11], DenseNet [8] and ResNet [9]. We use VGG19 and ResNet50 pre-trained on ImageNet to extract video embeddings, whose dimensions are 512 and 2048, respectively. Places365 is a image database proposed for scene recognition and the authors investigated the performance of different networks in [7], among which the pre-trained DenseNet161 and ResNet50 are used to learn deep features for the AVSC task with dimensions of 2208 and 2048. We also train the ResNeSt50 [10] model using the Places365 database by ourselves. These pre-trained ImageNet-CNNs and Places365-CNNs are used to extract deep visual features for multimodal scene classification.

2.2.3. Text Embedding

In our previous work [21], the acoustic segment model (ASM) approach can achieve a competitive performance in the ASC task. The main idea of AVSM approach is to represent a scene as a temporal sequence of fundamental units by using acoustic and visual features simultaneously, which is able to analyze the finer information in detail. The AVSM training procedure contains two stages. First, we use the Gaussian mixture model and hidden Markov model (GMM-HMM) to do initial segmentation. Then we use the CNN-HMM model to generate the final AVSM sequence based on initial segmentation. The AVSM sequence is translated into embedding through a text categorization method, which is usually composed of two parts: term frequency and inverse document frequency (TF-IDF) [22]. From this perspective, the generated embedding could be considered as text feature.

2.3. Two-stage Classification

We adopt a two-stage classification strategy to improve the performance of multimodal systems, which was proposed in our previous work [15]. A ten-classifier and a three-classifier are built by using audio and video data. The final scene class is predicted by score fusion of these two classifiers. More details can be found in [15].

Table 1: Pre-trained audio and video models used to extract embeddings.

Index	1	2	3	4	5
Audio Model	PANN	VGGish	FCNN	Resnet	
Video Model	ImageNet-VGG19	ImageNet-ResNet50	Places365-DenseNet	Places365-ResNet50	Places365-ResNeSt50

3. EXPERIMENTS AND RESULTS

3.1. Feature Extraction

The dataset for the AVSC task is TAU Urban Audio-Visual Scenes 2021 [23, 19]. The development dataset contains 34 hours of data with time-synchronized audio and video content. All audio and video samples have a length of 10 seconds, and we split them into 1-second samples. Based on the official training/test split, there are 86460 training data and 36450 test data.

To extract audio embedding, we use log-mel filter bank (LMFB) features with delta and delta-delta operations, which generates a feature shape of $39 \times 128 \times 6$. For video data, images are extracted from video clips and resized to 224×224 patches. We calculate the video embedding at a frame rate of 2 fps. To learn text embedding, the pre-trained VGG19 model is applied to extract visual feature. Mel-frequency cepstral coefficients (MFCC) is adopted as acoustic feature. We concatenate these two features in frame level as the input of AVSM training, and LMFB features are used to train FCNN-HMM.

3.2. Model Training

The classifier shown in Figure. 1 is a simple feed-forward neural network (FNN) which contains three layers of size 512, 128 and 64. The size of the output layer is 10 for ten-class classifier and 3 for three-classifier. Adam optimizer is used to train our models. Both pre-trained audio and video models are fine-tuned during training with a small learning rate of 0.00001. The weight decay is set to 0.00001 and the batch size is set to 32. All our models are trained using PyTorch toolkit.

In general, we train several DNN architectures with different combination of audio, video and text embeddings. Specifically, pre-trained audio and video models used to extract embeddings are listed in Table 1.

3.3. Results on Development Dataset

We first compare the performance of AVSC models when using different audio embeddings. Table 2 shows the performance comparison of three kinds of audio embeddings on the development dataset, where "A1-V1" denotes that the audio model is PANN and the video model is ImageNet-VGG19 as shown in Table 1. Please note that PANN and VGGish are pre-trained using AudioSet while FCNN is trained using the audio data of development dataset. In these preliminary experiments, data augmentation methods were not used and the video embedding was calculated at a frame rate of 1 fps. As shown in Table 2, all these three multimodal systems outperform the official baseline system. Moreover, FCNN trained with official data is able to provide more effective audio embedding. Therefore, in the following experiments, CNN-based models trained with official data are used to extract audio embedding.

In Table 3, we present experimental results on the development dataset for different multimodal systems. The audio and video models used to extract embeddings are denoted in Table 1. Systems shown in the left are trained with audio and video embeddings only while systems shown in the right are the corresponding versions fused with text embedding.

Table 2: Experimental results on the development dataset when using different audio embeddings.

System	Baseline	A1-V1	A2-V1	A3-V1
Log loss	0.658	0.521	0.470	0.329
Accuracy (%)	77.0	87.2	87.9	91.0

Firstly, the two-stage classification method can bring consistent improvement on both log-loss and accuracy metrics for all multimodal systems. For example for Idx.(11) with FCNN served as audio and ImageNet-ResNet50 served as video model, the log-loss and accuracy are 0.237 and 93.3%, respectively. By adopting the two-stage classification fusion, the log-loss is decreased by 21.1% and the accuracy is increased by 0.2%. Similar results are achieved for the AVTSC systems.

Secondly, besides audio and video embeddings, we investigate to analyze the scene with more finer information, which is called AVSM sequence. By comparing the models shown in the left and those in the right, we observe systems fused with text embedding yield better or comparable results than those without using text embedding. Especially for Idx.(5) where the audio model Places365-ResNeSt50 is trained by ourselves, the log-loss and accuracy are improved by 19.0% and 1.3%, respectively. When fusing with text embedding, our single multimodal system Idx.(18) achieves the best performance with a log-loss of 0.155 and an accuracy of 94.9%, which significantly outperforms the official baseline system. Compared with the official baseline system, the relative gains for log-loss and accuracy are 76.4% and 23.2%, respectively.

The bottom block of Table 3 shows our emsemble systems. Model ensemble is performed on two-stage classification system. As shown in Idx.(25) to Idx.(28), model ensemble can further improve the performance. The best ensemble system is obtained by fusing eight models as shown in Idx.(28) with a log-loss of 0.141 and an accuracy of 95.3%.

3.4. Submission Summary

In this section, we introduce our submitted systems. In our final submission, we use all data in the development dataset to train the systems. Besides the provided training/test fold, we also split the development dataset into three other training/test folds. We adopt four-fold cross-validation training strategy in our final submission. Because the research time is limited in the challenge, we only build AVTSC systems using data of fold 1 and fold 2. Our submitted systems are summarized as follows.

- Submission 1: Ensemble of four AVTSC models.
  These four AVTSC models are "A3-V3-T", "A3-V4-T", "A3-V5-T" and "A4-V3-T". They are all trained using data of fold 1 and fold 2.
- Submission 2: Ensemble of four AVTSC and AVSC models.

Table 3: Experimental results on the development dataset for different multimodal systems. "Two-stage" means using the two-stage classification strategy. "Y" means that we used the method. Systems named with "-T" means that text embedding was used during training.

Idx.	System	Two-stage	Log loss	Accuracy (%)	Idx.	System	Two-stage	Log loss	Accuracy (%)
(1)	A3-V3	-	0.206	94.2	(13)	A3-V3-T	-	0.202	93.9
(2)	A3-V3	Y	0.186	94.3	(14)	A3-V3-T	Y	0.190	93.9
(3)	A3-V4	-	0.213	93.7	(15)	A3-V4-T	-	0.209	93.9
(4)	A3-V4	Y	0.191	93.8	(16)	A3-V4-T	Y	0.184	94.0
(5)	A3-V5	-	0.216	93.6	(17)	A3-V5-T	-	0.175	94.8
(6)	A3-V5	Y	0.183	93.8	(18)	A3-V5-T	Y	0.155	94.9
(7)	A4-V3	-	0.234	93.6	(19)	A4-V3-T	-	0.215	93.9
(8)	A4-V3	Y	0.208	93.6	(20)	A4-V3-T	Y	0.190	93.9
(9)	A4-V4	-	0.229	93.6	(21)	A4-V4-T	-	0.243	93.6
(10)	A4-V4	Y	0.188	93.8	(22)	A4-V4-T	Y	0.210	94.1
(11)	A3-V2	-	0.237	93.3	(23)	A3-V2-T	-	0.248	92.8
(12)	A3-V2	Y	0.187	93.5	(24)	A3-V2-T	Y	0.199	93.5
(25)	Ensemble of {(6),(18)}						Y	0.147	94.7
(26)	Ensemble of {(6),(18),(12),(24)}						Y	0.145	95.1
(27)	Ensemble of {(2),(14),(6),(18),(12),(24)}						Y	0.143	95.2
(28)	Ensemble of {(2),(14),(6),(18),(10),(22),(12),(24)}						Y	0.141	95.3
(29)	Official Baseline						-	0.658	77.0

Four AVTSC models are "A3-V3-T", "A3-V4-T", "A3-V5-T" an "A4-V3-T". They are all trained using data of fold 1. Four AVSC models are "A3-V3", "A3-V4", "A3-V5" an "A4-V3". They are all trained using data of fold 2.

- Submission 3: Ensemble of four AVSC models.
  These four AVSC models are "A3-V3", "A3-V4", "A3-V5" and "A4-V3". They are all trained using four-fold cross-validation strategy.
- Submission 4: Ensemble of six AVSC models.
  These six AVSC models are "A3-V3", "A3-V4", "A3-V5", "A4-V3", "A4-V4" and "A3-V2". They are all trained using four-fold cross-validation strategy.

4. CONCLUSION

In this technical report, we propose an ensemble approach to solve the AVSC task in DCASE 2021 Challenge. To exploit effective deep features, we employ several pre-trained audio and video models to learn audio and video embeddings. Audio embedding extracted from models trained with official data is effective and video embedding extracted from models pre-trained on image datasets is effective. Moreover, we propose to use acoustic and visual segment model to represent finer information in a scene and then extract complimentary textual feature. By using early fusion method, audio, video and text embeddings are concatenated and fed to the classifier. A two-stage classification fusion is used to boost the performance of multimodal systems. We have shown that by combining effective audio, video and text embeddings, the performance of the AVSC task can be greatly improved. With the model ensemble approach, our system can achieve a log-loss of 0.141 and an accuracy of 95.3% on the development dataset, yielding 78.6% and 23.8% relative gains compared to the official baseline system.

5. REFERENCES

[1] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 9–13. [Online]. Available: https://arxiv.org/abs/1807.09840

[2] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, 2020, submitted. [Online]. Available: https://arxiv.org/abs/2005.14623

[3] S. Suh, S. Park, Y. Jeong, and T. Lee, "Designing acoustic scene classification models with CNN variants," DCASE2020 Challenge, Tech. Rep., June 2020.

[4] H. Hu, C.-H. H. Yang, X. Xia, X. Bai, X. Tang, Y. Wang, S. Niu, L. Chai, J. Li, H. Zhu, F. Bao, Y. Zhao, S. M. Siniscalchi, Y. Wang, J. Du, and C.-H. Lee, "Device-robust acoustic scene classification based on two-stage categorization and data augmentation," DCASE2020 Challenge, Tech. Rep., June 2020.

[5] W. Gao and M. McDonnell, "Acoustic scene classification using deep residual networks with focal loss and mild domain adaptation," DCASE2020 Challenge, Tech. Rep., June 2020.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[7] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[8] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[10] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, *et al.*, "Resnest: Split-attention networks," *arXiv preprint arXiv:2004.08955*, 2020.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[12] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[13] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[14] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[15] H. Hu, C.-H. H. Yang, X. Xia, X. Bai, X. Tang, Y. Wang, S. Niu, L. Chai, J. Li, H. Zhu, *et al.*, "A two-stage approach to device-robust acoustic scene classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 845–849.

[16] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[17] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[18] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *arXiv preprint arXiv:1805.09501*, 2018.

[19] S. Wang, A. Mesaros, T. Heittola, and T. Virtanen, "A curated dataset of urban scenes for audio-visual scene analysis," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, accepted. [Online]. Available: https://arxiv.org/abs/2011.00030

[20] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3852–3856.

[21] X. Bai, J. Du, Z.-R. Wang, and C.-H. Lee, "A hybrid approach to acoustic scene classification based on universal acoustic models." in *Interspeech*, 2019, pp. 3619–3623.

[22] D. Hull, "Improving text retrieval for the routing problem using latent semantic indexing," in *SIGIR94*. Springer, 1994, pp. 282–291.

[23] S. Wang, T. Heittola, A. Mesaros, and T. Virtanen, "Audio-visual scene classification: analysis of dcase 2021 challenge submissions," 2021.